Data Enrichment and Cleansing in Banking Applications

Muhammad Ali[1], Engr. Muhammad Anwar Usman[2]

[1]*MSCS SZABIST, Karachi*
Ali_ck@hotmail.com

[2]*Perfect Engineering, Karachi*
anwarusman@szabist.edu.pk

Abstract: *More over all banks in Pakistan are involved in data inconsistency problem in a such a way that the information of customers they have over 40 50 years are not complete or not updated .There for every bank is moving towards a centralized database or a core banking system where they can have a central information for all customers. To do so when the banks migrated millions of accounts from branch to its centralized server over the years, incomplete and incorrect information migrated over too. Also, new fields were added in the system like computerized national identity card which is mandatory now for the State Bank of Pakistan to have for each account which is not updated for existing users over the years.*

Keywords: *Data Cleaning, Account Opening, Banks.*

1. INTRODUCTION

More over all banks in Pakistan are involved in data inconsistency problem in a such a way that the information of customers they have over 40 50 years are not complete or not updated .There for every bank is moving towards a centralized database or a core banking system where they can have a central information for all customers.

To do so when the banks migrated millions of accounts from branch to its centralized server over the years, incomplete and incorrect information migrated over too.

Also, new fields were added in the system like computerized national identity card which is mandatory now for the State Bank of Pakistan to have for each account which is not updated for existing users over the years.

1.2 Drawbacks for Inconsistent Data

Banks wants to introduce new services but they could never keep the track of count that how many customers uses which services in a consistent fashion.

Strategic decisions could not be made due to incorrect account information.

Requirement of a State Bank of K.Y.C (know your customer).

1.3 Limitations

As the branches are spread all over the Pakistan in urban and rural areas, they are not connected to central server by any means cause of lack of high speed internet connection .So many branches run in the disconnected mode and transfer the data at the end of particular date/time.

2. REASONS FOR INCORRECT DATA

2.1 Application Errors

Sometimes the Software running at branch side is unable to verify and validate the correct data that has been entered into the system [2]

2.2 Human Error

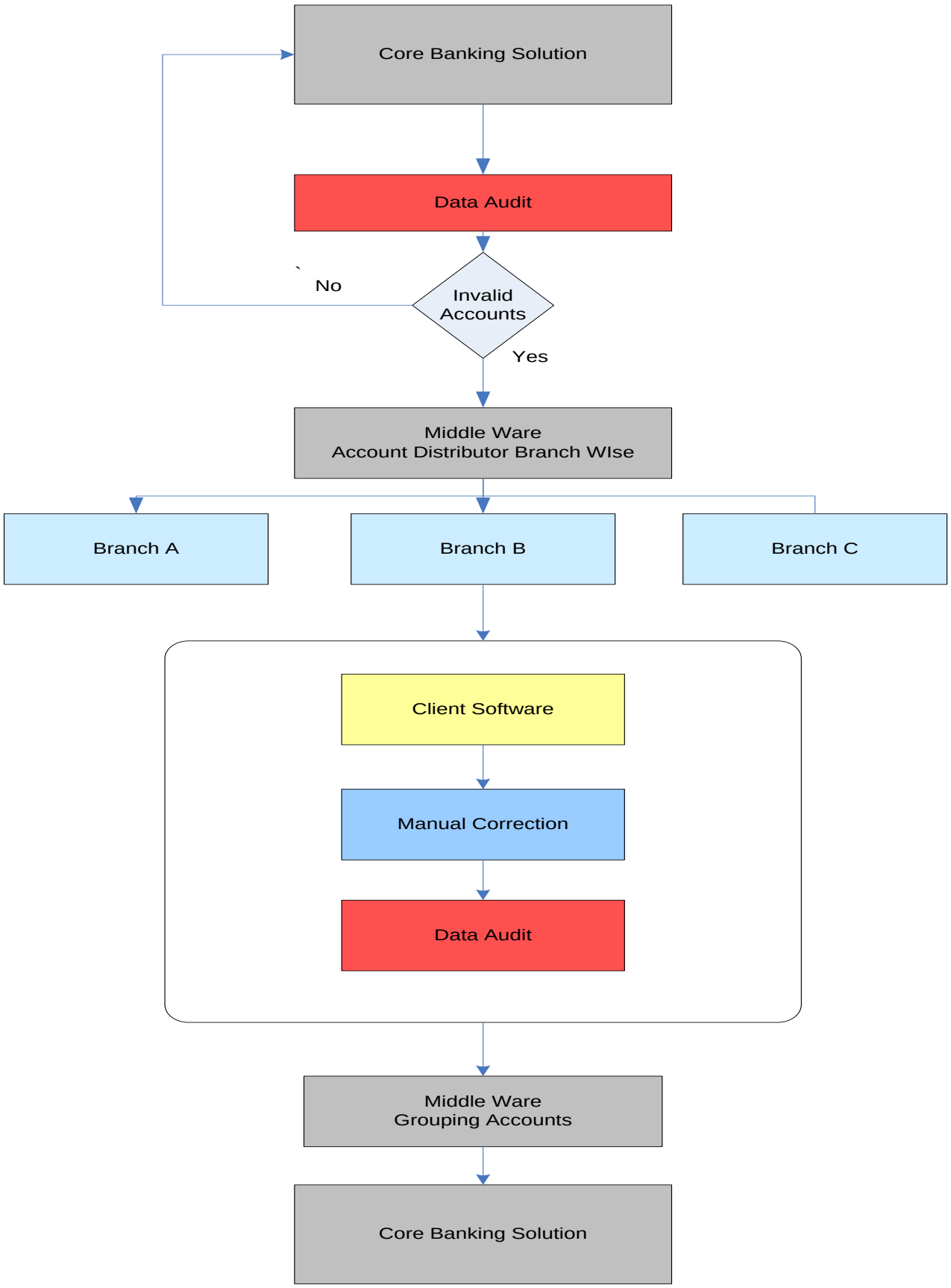
This is the main cause of invalid data in the system. Sometimes the end user who is using the system enters incorrect or miss information.

But if suppose these are validated by the software but software's cant validate logical errors .For instance if the end user at branch end enters account opening date in the closing date column then incorrect information will be entered into the system though that data type was correct for the system. [2]

2.3 Deliberate Manipulations

Branch User must be forced to enter that is wrong but is predictable because the system would reject the data otherwise. [2]

3. METHODOLOGY



3.1 Core Banking Solution

Core Banking Solution refers to the a set of a complete solution where all banking transactions are maintained along with all the data that is relevant to Retail banking, corporate/wholesale banking which drive revenue generation, performance and efficiency improvements for financial institutions across the globe.

Here core banking Solution is meant by a repository which contains all the data relevant to Account Opening of a customer.

3.2 Data Audit

This Step will again check the data that is corrected by branch user and will generate the summary of records that is corrected and that cannot be corrected due to some reasons that are

1. Fake NIC
2. No NIC
3. Cannot reach to customer on given address/phone[2].

3.3 Reports

It is essential to confirm the invalid or missing data that is identified and extracted by the error identification process (Data Audit).. This verification is done through crystal reports of dirty data based on error type.

3.4 Distributor

This middle ware is the main software at the central side which splits the account data according to branch wise in some file format (txt, XML). So that all the related invalid accounts are stored in the file; which are related to that particular branch. Then the file is sent electronically through mail to their respected branches.

3.5 Branch Client Software

Branches have client software that reads the file and displays the information on a form where he/she can edit the data. The branch user then fills up the information that is missing or invalid. For new fields those are not present like Computerized National Identity Card, the branch user then calls/email/mail the customer and takes valid values.

3.6 Manual Correction

Manual correction is the main part of data cleansing where data is checked and cleansed through client software that will display the data on the screen in a form which will and branch user will verify it through his system

3.7 Data Audit

This Step will again check the data that is corrected by branch user and will generate the summary of records that is corrected and that cannot be corrected due to some reasons that are

1. Fake NIC
2. No NIC
3. Cannot reach to customer on given address/phone.

3.8 Middle Ware (For Grouping Accounts)

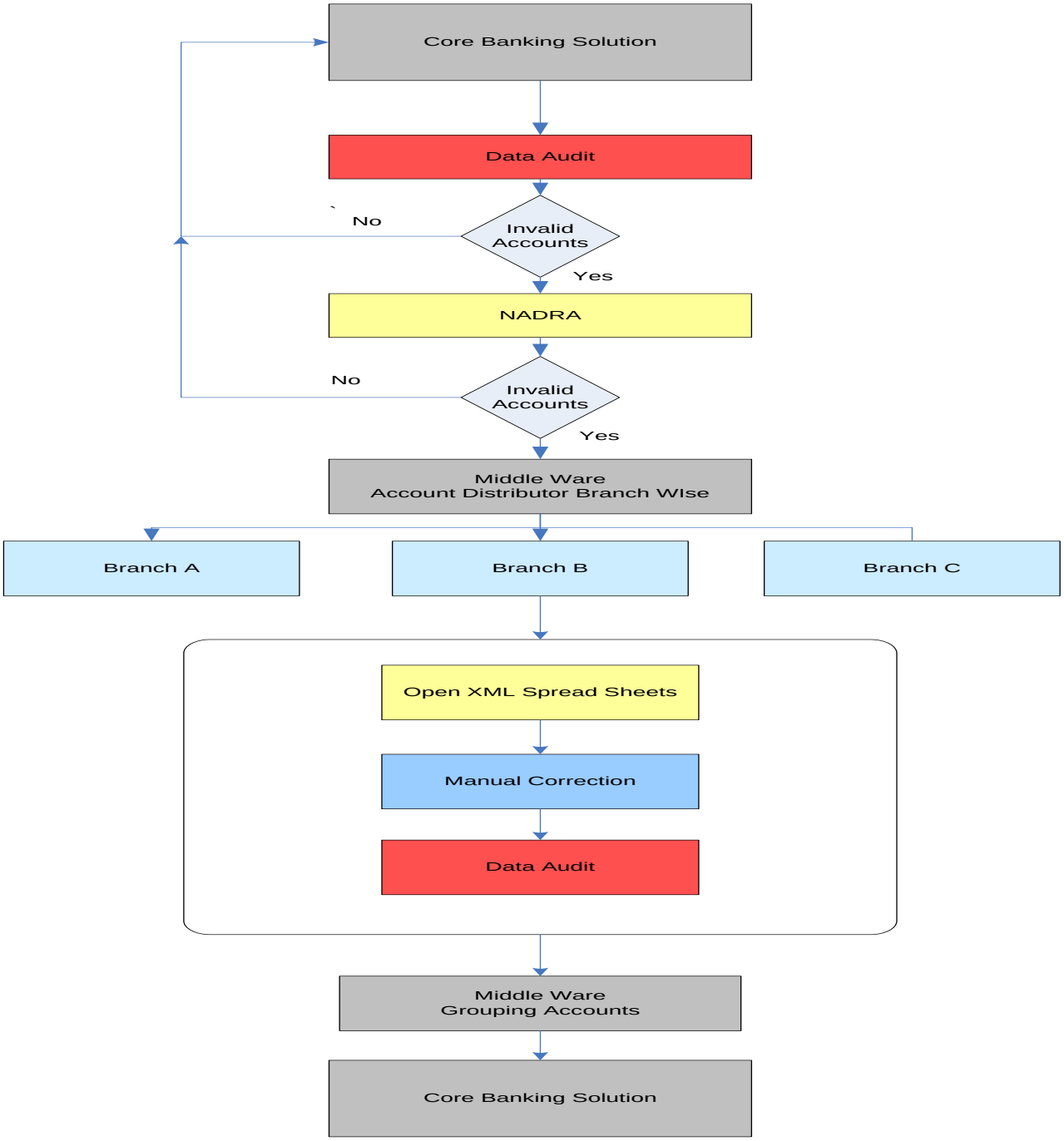
In this step the validated/corrected data from various branches group together and then a file is generated that is restored the in the central repository which can later restore to banking solution.

4. PROPOSED MODEL

4.1 Drawbacks of the Methodology

- Manual checking and verification that is time consuming.
- Branches have different operating systems and for that different client modules are developed according to operating system
- Fake NIC are not verified
- Client side software needs to be updated every time as new fields are added to the system

4.2 Proposed Approach



This approach has two main modifications

4.2.1 NADRA

This is one of the major changes in the current structure. Invalid accounts that have a NIC or CNIC are checked through Nadra for further details .In this way we can get CNIC from NIC if the account is not updated .Also the fake NIC accounts can be verified from NADRA .If NIC does not exist then account is subject to terminate .Accounts that does not have NIC will go for further cleaning on Branch side where branch user will complete it details through email/mail or if not found terminate the account .In this way lot of time is saved where branch user verifies the account through manual checking.

4.2.2 OPEN XML SPREADSHEETS

This is another effective change in the process where branches does not have to install client software to update

account information .spreadsheet is directly created by middleware and sent to branches ,branches then update the data and send the spreadsheet directly to middleware to group the accounts[5]. In this way lot of time is saved .One major advantage is that if a new field is added to the system, client software at all branches does not need to be updated, rather have to just fill/update the spreadsheet that is independent of operating system running at branch side.

5. CONCLUSION AND FUTURE DIRECTION

The conclusion of this IS report gives a working way of data cleansing in banks of Pakistan regarding account opening problem. I have a provided a solution for some kind of automation in the process through NADRA and Open XML. The future direction will be a way to automate the process fully through any means which will save time and effort in a better way

REFERENCES

[1].Wise Geek. Intern: http://www.wisegeek.com/what-is-data-cleansing.htm

[2].White Paper. 'Oracle and Trillium Software'.

[3]. Review. Internet: http://www.dmreview.com/dmdirect/20041029/1012952-1.html

[4]White Paper on 'Office Open XML'. EMCA.

[5]. Wikipedia. Internet: http://en.wikipedia.org/wiki/Office_Open_XML